A ferramenta *FreP* e a frequência de tipos silábicos e classes de segmentos no Português

Marina Vigário[1], Fernando Martins[2,3] e Sónia Frota[2]
[1]Universidade do Minho, [2]Universidade de Lisboa, [3*]IUTEC

0. Introdução

A importância da informação de frequência para efeitos da compreensão do comportamento linguístico dos falantes, bem como do próprio processo de aquisição do sistema gramatical, tem sido mostrada por uma multiplicidade de trabalhos recentemente publicados (e.g. Booij, 1995, 2004; Thornton, 1996; Beckman e Edwards, 2000; Roark e Demuth, 2000; Bybee e Hopper, 2001; Jurafsky, Bell e Girand, 2002; Vigário, 2003; Demuth e Johnson, 2003; Prieto, 2004), como já salientado em Vigário, Martins e Frota (2005). Cientes disso e face à dificuldade em conhecer-se dados de frequência sobre unidades fonológicas no Português, em contextos diversificados, concebemos e implementámos um projecto visando a criação de uma ferramenta electrónica capaz de extrair, a partir de texto escrito, informação sobre a frequência de ocorrência de unidades fonológicas, bem como a exploração dessa informação em vários domínios da análise linguística. Este projecto, em conjunto com uma versão inicial da ferramenta, que designámos *FreP*, foi apresentado na edição anterior do *Encontro Nacional da Associação Portuguesa de Linguística* (cf. Vigário, Martins e Frota, 2005).

O presente trabalho sintetiza as possibilidades da versão anterior da ferramenta e os resultados que puderam já ser obtidos com o recurso a ela. Para além disso, mostra os avanços entretanto feitos. Tal como anteriormente, para além das novas funcionalidades do *FreP*, apresentam-se os resultados obtidos através dessas funcionalidades, e sugerem-se eventuais consequências para a análise linguística. Como ficará claro, a ferramenta tornou-se mais poderosa ao permitir extrair, para além de informação de frequência sobre a palavra prosódica e os clíticos, também informação relativa à sílaba e aos segmentos, assim como à localização do acento de palavra. Para além disso, a aplicação é agora de utilização mais agradável e acessível, tendo sido feito um esforço no domínio da sua aparência visual e da transparência na sua utilização. Finalmente e tendo em conta o trabalho até aqui já desenvolvido, projecta-se as direcções futuras, que passarão, proximamente, por dotar o *FreP* da capacidade de extrair informação sobre a frequência de *todas* as unidades fonológicas de nível idêntico ou inferior ao da palavra prosódica.

XXI Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2006, pp. 675-687

1. Versão anterior do *FreP* e resultados já obtidos

Nesta secção lista-se as funcionalidades da versão anterior do *FreP* e resume-se algumas das implicações que os resultados obtidos através da sua aplicação já tiveram em relação a diversos aspectos da investigação linguística.

1.1. Breve descrição das funcionalidades da versão anterior do FreP

Na versão anterior, o *FreP* permitia (i) localizar vogais, (ii) dividir sílabas, (iii) determinar a presença e a localização do acento de palavra. Em função desta informação básica, um conjunto de passos possibilitavam extrair outros tipos de dados: (i) número de sílabas por palavra; (ii) número de palavras com uma, duas, três, ... N, sílabas; (iii) número de palavras prosódicas; (iv) número de palavras clíticas; (v) número de palavras prosódicas com uma, duas, três, ... N, sílabas; (vi) número de palavras clíticas com uma ou duas sílabas; (vii) número de palavras prosódicas monomoraicas (e não-monomoraicas), termo usado para designar palavras terminadas em vogal oral; (viii) número de palavras clíticas monomoraicas (e não-monomoraicas).

1.2. Resultados anteriores obtidos com o recurso ao *FreP*

Dadas as suas potencialidades, a versão anterior permitiu já avançar em três áreas da investigação linguística.

Mostrou-se que os dados de frequência favorecem a hipótese de que o Português (Europeu) não é sensível à restrição de *palavra mínima* (Vigário, Martins e Frota, 2005). Efectivamente, apesar de, no léxico desta língua, não existir um número elevado de palavras monossilábicas terminadas em vogal (que classificámos como *monomoraicas* para efeitos de comparabilidade com o que vem sendo descrito para outras línguas), a frequência de ocorrência das palavras com este formato é bastante elevada.

Sobre este tópico, devemos agora introduzir um parêntesis. Tendo sido detectado um erro na fórmula de cálculo da versão inicial do *FreP* relativa a este tipo de palavras (ver detalhes na secção seguinte), aproveita-se para fazer aqui uma correcção aos dados apresentados anteriormente em Vigário, Martins e Frota (2005: Quadro 1, p. 903): onde se lê 11,66 (percentagem de palavras monossilábicas com sílaba fechada), deve ler-se 24,06 e onde se lê 19,8 (percentagem de palavras monomoraicas), deve ler-se 7,4. Embora os valores anteriores e os corrigidos sejam muito distintos, a conclusão a que se chega não é alterada, uma vez que o valor agora obtido para as palavras classificadas como monomoraicas se mantém suficientemente elevado, sendo mesmo similar ao apresentado para palavras com quatro ou mais sílabas. Note-se que este número não pode ser comparado em termos absolutos com os apresentados para palavras com duas e mais sílabas, dado que ele é relativo a *dois* parâmetros (número de sílabas por palavra *e* constituição da sílaba), enquanto os restantes são relativos *apenas* ao número de sílabas por palavra, independentemente da sua constituição.

O recurso ao *FreP* produziu um segundo conjunto de resultados, agora no domínio da aquisição da linguagem (Vigário, Freitas e Frota, no prelo). Especificamente,

estudou-se a (ordem de) emergência de palavras com diferentes formatos (tamanho, medido em número de sílabas) nas primeiras produções linguísticas das crianças, e mostrou-se como ela se relaciona positivamente com o predito em função da frequência de ocorrência dos diferentes formatos de palavra na fala adulta (entre adultos) e na fala dirigida à criança.

Um terceiro tópico investigado com o recurso ao *FreP* permitiu avaliar a importância relativa dos dados de frequência face aos princípios ou regras gramaticais relevantes para a colocação dos clíticos pronominais (Vigário, Martins e Frota, 2005). Concretamente, foi possível mostrar que, embora os proclíticos fonológicos sejam muitíssimo mais frequentes do que os enclíticos fonológicos, esse facto não parece inibir a tendência inovadora recente para uma colocação pós-verbal generalizada dos clíticos pronominais, originadora de ênclise fonológica e independente dos contextos sintácticos.

2. O *FreP* – Nova versão

A nova versão do *FreP* encontra-se melhorada de vários pontos de vista. Esta secção sintetiza o essencial das novidades introduzidas.

2.1. Novas funcionalidades

Tal como projectado anteriormente, a nova versão do *FreP* permite agora as seguintes operações (para além das já antes disponibilizadas):

- Identificar e contar os tipos silábicos (CV, V, CVC, ...), e isto (i) em função da posição na palavra (inicial, medial, final e em palavras monossilábicas); (ii) em função das sílabas serem ou não acentuadas; e (iii) em função da posição na palavra *e* do acento;
- Identificar e contar os segmentos pertencentes às grandes classes C, V e G, assim como a ocorrência do traço nasal em segmentos não-consonânticos e de posições vocálicas (*V-Slots*) entre sequências consonânticas indutoras de violações de princípios de silabificação (e.g. *optar* > o.pV.tar);
- Identificar e contar o número de palavras (prosódicas) com mais de uma sílaba com acento final, penúltimo e antepenúltimo (anteriormente o programa permitia apenas distinguir entre palavras com acento e palavras sem acento, assim se obtendo a separação entre *palavras prosódicas* e *clíticos*);
- Contabilizar o número total de palavras ortográficas;
- Contabilizar o número total de caracteres ortográficos.

2.2. Critérios de segmentação, identificação ou categorização das unidades

Para a segmentação, identificação ou categorização das unidades fonológicas consideradas na versão actual do programa, foram seguidos os seguintes critérios fundamentais:

- Sobre os tipos silábicos (CV, V, CVC, ...) – para efeitos da contagem de tipos silábicos, entendeu-se como *diferente tipo silábico* aquele que apresenta uma diferente composição/organização em termos de consoantes (C), vogais (V), glides (G) e traço nasal em Rima (N); neste quadro, importa lembrar que as únicas origens de sílabas contendo ditongos crescentes são as resultantes de glides ambissilábicas e de glides pós-tónicas, obrigatórias (cf. Vigário, Martins e Frota 2005);
- Sobre a posição dos tipos silábicos – aqui deve clarificar-se que a noção de *posição* é relativa a unidades que podem ser classificadas como palavras prosódicas e/ou morfossintácticas, ou, dito de outro modo, palavras prosódicas ou clíticos;
- Sobre as grandes classes de segmentos (C, V, G) – a este respeito, lembramos que, para todas as operações com o *FreP* e tal como já foi dito em Vigário, Martins e Frota (2005), se consideram na classe das Gs apenas as glides obrigatórias, isto é, as pertencentes a ditongos decrescentes obrigatórios (e.g. *pau*) e a ditongos crescentes pós-tónicos, em que também não é possível a glide alternar com vogal (e.g. *família*); de igual modo se lembra que se assume que [k^w] e [g^w], em palavras como *quando* e *guarda*, são oclusivas labializadas e não sequências de oclusiva e semivogal (cf. Andrade e Viana, 1994; Vigário e Falé, 1994; Viana *et al.*, 1996); finalmente, no interior da classe V estão todas as vogais, independentemente de serem orais ou nasais;
- Sobre o traço nasal em segmentos não-consonânticos – optou-se por fornecer informação separada sobre a ocorrência do traço nasal (ou autossegmento nasal) responsável pela nasalização de vogais e semivogais em Rima; para esta contabilização, assumiu-se o que se crê ser a representação fonológica da nasalidade (cf. Andrade e Kihm, 1988; Vigário, 2003: Cap. 3); em conformidade com isto mesmo, considerou-se apenas uma ocorrência do traço em ditongos nasais;
- Sobre as posições vocálicas (*V-Slots*) – entendeu-se ser útil distinguir entre vogais fonológicas e posições vocálicas entre sequências consonânticas indutoras da violação de princípios de silabificação, que são susceptíveis de ser preenchidas por vogais epentéticas (e.g. *optar* > o.pV.tar) (cf. Vigário e Falé, 1994; Mateus e Andrade, 2000); designámos essas posições como *V-Slots*;
- Sobre o número total de palavras ortográficas – de acordo com a acepção convencional de *palavra ortográfica*, este número é obtido a partir da contagem de uma ou várias letras contíguas precedidas e seguidas por espaço branco ou sinal de pontuação.

Importa, finalmente, lembrar que a ferramenta se encontra no presente momento optimizada para o Português Europeu.

2.3. Correcções à versão anterior e procedimentos minimizadores da ocorrência de erros

Foram introduzidas algumas correcções/alterações em relação às funcionalidades já disponíveis na versão anterior, sendo as mais importantes as que passamos a apresentar.

Como vimos acima, foi detectado um erro na fórmula de contagem de palavras monomoraicas (isto é, palavras com uma única sílaba e terminadas em vogal oral). Contrariamente ao que era relevante para esta classe de palavras, estavam nessa versão a ser, incorrectamente, incluídas nesta classe palavras terminadas em glide. A presente versão está já corrigida a este respeito.

Dada a natureza desta ferramenta, existe um conjunto de erros persistentes, que decorrem da impossibilidade de se estabelecerem correlações perfeitas entre a representação ortográfica e as unidades fonológicas extraídas. Nestes casos, apenas a listagem das excepções ou a introdução de procedimentos de intervenção nos ficheiros de texto originais, antes de correr o programa, permitem minimizar os erros. Em três classes de casos, foram introduzidos no *FreP* procedimentos automáticos visando excluir da classe de erros relevante um pequeno conjunto de palavras específicas, dada a sua elevada frequência. Fala-se a seguir em cada uma.

Um erro que inevitavelmente persistirá correndo o *FreP* diz respeito à atribuição de acento à primeira palavra prosódica de palavras formadas com os sufixos *–mente* e *-z-avaliativos*, se ela for monossílábica e terminar em vogal (e.g. *mamente*). Para minimizar este tipo de erro introduziu-se um procedimento automático que permite agora excluir deste conjunto de erros as palavras muito frequentes *somente* e *sozinho/a(s)*.

Dada a sua fórmula para a localização do acento, o *FreP* não identifica o acento no local apropriado na primeira palavra prosódica de palavras formadas com os sufixos *–mente* e *-z-avaliativos*, se ela apresentar um padrão acentual irregular face às regras de acentuação ortográfica (e.g. *agilmente*). Para minimizar este tipo de erro, introduziu-se um procedimento automático que permite agora excluir deste conjunto de erros as palavras muito frequentes *dificilmente* e *facilmente*.

Finalmente, uma outra classe de erros, decorrente da ausência de uma relação sistemática entre o sistema de representação ortográfica e as unidades fonológicas, diz respeito às palavras cuja terminação coincide com a dos sufixos com acentuação independente da base morfológica mas que não são formadas com estes sufixos (e.g. *arrozinho*). Para minimizar este tipo de erro, introduziu-se um procedimento automático que permite agora excluir deste conjunto de erros as palavras frequentes *vizinho/a(s)*, *cozinha(s)*.

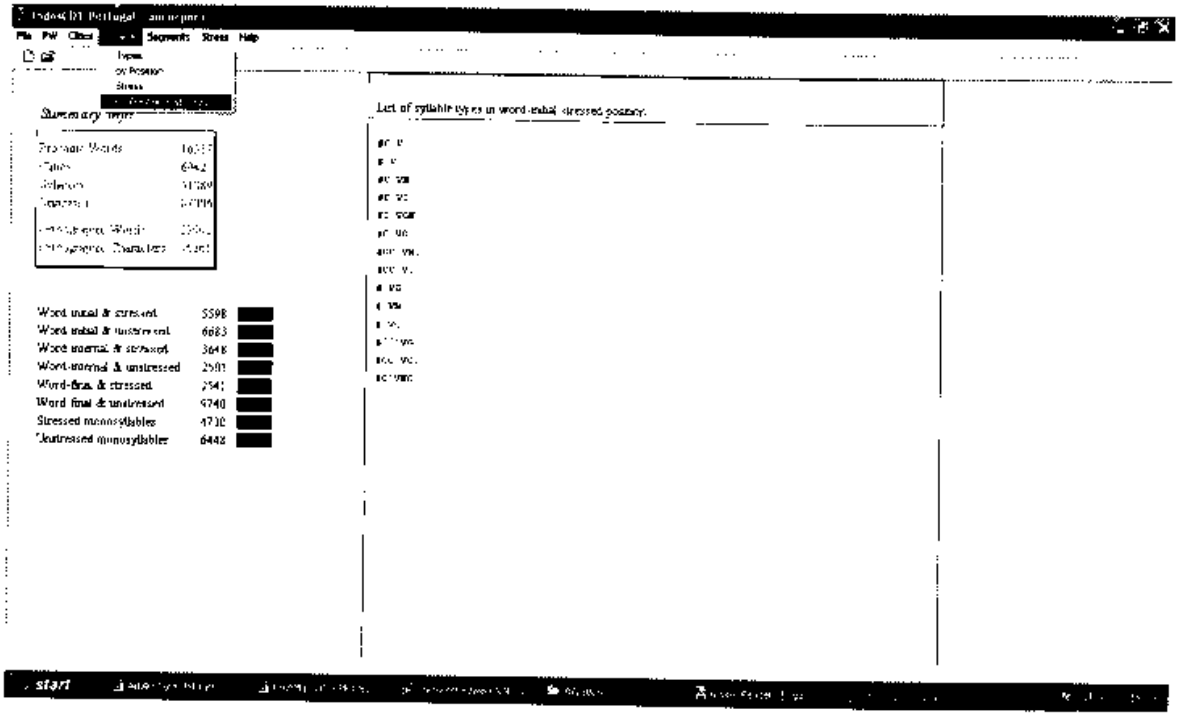
2.4. Manual

Entre as melhorias mais significativas introduzidas nesta versão do *FreP*, conta-se a disponibilização de um *Manual*, onde são detalhadas informações diversas relativas ao programa. O *Manual* da ferramenta inclui: uma apresentação do *FreP*, contendo uma caracterização geral da aplicação; notas sobre a génese e desenvolvimento deste projecto; como adquirir ou actualizar o programa; os requisitos de utilização, incluindo especificação do tipo de ficheiro em que pode correr; instruções para a instalação do programa e como carregar um ficheiro; as funcionalidades da ferramenta, com a identificação dos comandos relevantes e descrição das funções por eles executadas; os

critérios adoptados para a segmentação das diversas unidades fonológicas; as opções disponíveis (ou a disponibilizar futuramente) ao utilizador; uma listagem das limitações do programa, com sugestões de procedimentos para as contornar; e, finalmente, indicações dos trabalhos já disponíveis feitos com o recurso ao *FreP*.

2.5. Aspecto visual

Um domínio no qual se fez um grande investimento na presente versão diz repeito à aparência da ferramenta. Procurou-se torná-la visualmente atraente e dispor a informação do modo mais transparente. Com este intuito, optou-se pelo formato *windows*, com um sistema de comandos e janelas conhecido da generalidade dos utilizadores de programas informáticos, na sociedade em geral e na comunidade científica da área, em particular. A figura seguinte mostra o aspecto do ecrã depois de aberto o *menu Syllable*, seleccionada a opção *by Position and Stress* e pedida a listagem de sílabas acentuadas em posição inicial de palavra (prosódica).



2.5. Disponibilização da informação

Foi criado um sítio electrónico com informações várias relativas ao *FreP*, cujo endereço é http://www.fl.ul.pt/LaboratorioFonetica/FreP.

Nessa página pode encontrar-se uma breve descrição do programa, vídeos de demonstração do seu funcionamento, o *Manual* do programa, publicações e apresentações sobre ou que exploram o *FreP*, em formato .pdf e susceptíveis de *download*, bem como a forma de obter ou actualizar o programa.

3. Resultados novos

Apresentam-se aqui alguns dos resultados obtidos a partir das novas funcionalidade dos *FreP*. Para tal, recorreu-se ao *corpus* anteriormente manipulado e já descrito em Vigário, Martins e Frota (2005), aí designado por TA90PE. Lembra-se que se trata de uma amostra do *corpus* do *Português Falado. Documentos Autênticos*, editado em CR-ROM pelo Centro de Linguística da Universidade de Lisboa e Instituto Camões, que inclui os dados do Português de Portugal da década de 90 (CD 1).

Deve referir-se que os materiais foram tratados de modo a eliminar os cabeçalhos de cada entrevista que compõe esse *corpus*, dado este tipo de texto não fazer parte dos dados a tratar, nem seguir as convenções ortográficas do Português. Foram igualmente eliminadas as consoantes mudas presentes no texto. Para este último efeito, seguiu-se um procedimento semiautomático, descrito no Manual do programa.

Os resultados novos distibuem-se pelos seguintes quatro tópicos: (1) frequência relativa dos diferentes tipos silábicos na fala adulta; (2) frequência dos tipos silábicos no *input* e sua ordem de emergência nas primeiras produções; (3) frequência relativa das grandes classes de segmentos; e (4) frequência relativa das palavras com diferentes distribuições acentuais. O estudo respeitante à aquisição dos tipos silábicos foi apresentado autonomamente, pelo que se resume apenas os resultados aí obtidos e se remete para Frota, Freitas, Vigário e Martins (2005) e Freitas, Frota, Vigário e Martins (2005) para os detalhes sobre essa investigação. A apresentação do conjunto de resultados acima referido consitui o corpo das secções que se seguem.

Importa, para concluir esta introdução, chamar a atenção para o facto de estarmos ainda a trabalhar com uma versão do programa não integralmente testada. Se bem que no essencial se tenha confiança nos dados disponibilizados neste momento, não foi ainda possível fazer-se um trabalho sistemático e exaustivo de avaliação de cada funcionalidade. Este aspecto deve ser tomado em consideração na leitura dos dados apresentados.

3.1. Resultados novos (1): frequência relativa dos diferentes tipos silábicos na fala adulta

O primeiro conjunto de resultados que apresentamos diz respeito à frequência relativa dos diferentes tipos silábicos no presente *corpus* (ver Quadro 1).

Tipos	%	Tipos	%
CV	46,36	CVGC	1,21
V	15,83	VGN	0,59
CVC	11,01	CCVN	0,47
CVGN	5,62	CCVC	0,38
CVN	5,37	CGV	0,25
VC	3,03	CVGNC	0,17
CVG	2,66	GV	0,13
VN	2,64	CGVC	0,12
CCV	2,18	GVGN	0,12
VG	1,51	10 outros	< 0,10

Quadro 1: Frequência relativa dos diferentes tipos silábicos presentes no *corpus* TA90PE, num total de 41889 sílabas extraídas (C=Consoante; V=Vogal; G=Glide; N=autossegmento nasal).

Como seria de esperar face ao que conhecemos da literatura, onde dados semelhantes são apresentados para outros *corpora* (cf. Andrade e Viana, 1994; Vigário e Falé, 1994; Viana *et al.* 1996), o tipo CV é de longe o mais frequente, sendo o tipo seguinte mais frequente o constituído por apenas uma vogal. Estes, juntamente com os tipos CVC, CVN e CVGN, são os únicos que ultrapassam os 5% de frequência, sendo todos os restantes tipos silábicos pouco frequentes.

Uma observação dos dados tendo em conta a posição na palavra mostra que a ocorrência dos diversos tipos silábicos não se distribui, contudo, uniformemente por todas as posições consideradas (ver Quadros 2 e 3). Existe maior diversidade (e complexidade silábica) em posição final de palavra, em palavras monossilábicas e em posição inicial de palavra (Quadro 2).

Tipos silábicos diferentes por posição na palavra			
Inicial	Interna	Final	Monossílábica
15	8	23	16

Quadro 2: Número total de tipos silábicos diferentes por posição na palavra presentes no *corpus* TA90PE (em palavras com mais de uma sílaba e em palavras monossilábicas).

Ao contrário do tipo silábico CV, que apresenta uma distribuição mais equitativa pelas diferentes posições na palavra, a maioria das ocorrências do tipo silábico V concentra-se na posição inicial de palavra e nas palavras monossilábicas. Também os tipos silábicos CVC, CVGN, CVN, VN, CCV, VG e CVGC não apresentam uma distribuição homogénea: (i) as sílabas com N são mais frequentes em poisção inicial ou final e em palavras monossilábicas; (ii) as sílabas com G são mais frequentes em palavras monossilábicas; (iii) o tipo CVC encontra-se fundamentalmente em posição final de palavra; e (iv) o tipo CCV encontra-se predominantemente em posição inicial de palavra (Quadro 3).

Tipos	Inicial	Interna	Final	Monossílábica
CV	11,56	10,95	16,46	7,38
V	**6,58**	0,54	1,03	**7,68**
CVC	2,52	0,47	**5,88**	2,14
CVGN	0,66	0,00	**2,29**	**2,67**
CVN	**2,57**	1,37	0,42	1,01
VC	1,48	0,00	0,55	0,99
CVG	0,87	0,45	0,52	0,82
VN	**1,12**	0,20	0,00	**1,31**
CCV	**1,04**	0,62	0,51	0,00
VG	0,35	0,00	0,04	**1,13**
CVGC	0,00	0,00	0,29	**0,92**

Quadro 3: Frequência relativa dos diferentes tipos silábicos com frequência superior a 1% presentes no *corpus* TA90PE em função da posição na palavra (Inicial, Interna, Final e em palavras monossilábicas).

Se se tiver em conta que as margens da palavra podem constituir posições proeminentes, tipos silábicos relativamente pouco frequentes mas que se encontram nessas posições poderão assumir maior destaque na fonologia da língua do que outros tipos silábicos igualmente pouco frequentes.

Uma constatação semelhante se pode fazer quando considerada a ocorrência de tipos silábicos em função da presença ou ausência de acento de palavra (ver Quadro 4), informação, até onde sabemos, previamente inexistente na literatura. Em termos relativos, o tipo CV ocorre muito mais frequentemente em posição átona do que em posição tónica, o mesmo sucedendo com o tipo V, e ainda CVC, VC e CCV. Estes factos resultam, naturalmente, de haver uma maior proporção de sílabas átonas (60.54% do total de sílabas), relativamente às tónicas (39.36%), o que, por sua vez decorre de a maior parte das palavras da língua ser di e polissilábica (daí resultando, portanto, um maior número de sílabas átonas do que tónicas), bem como de haver mais palavras monossilábicas átonas do que tónicas. Neste contexto, será significativa a preponderância de sílabas tónicas para os tipos silábicos CVGN, CVN, CVG, VN, VG e CVGC. Por outras palavras, as sílabas com nasal e glide surgem maioritariamente em posição tónica, uma posição naturalmente proeminente, ao contrário das sílabas fechadas por C.

Tipos	Tónica	Átona
CV	12,71	**33,65**
V	5,77	**10,06**
CVC	4,40	**6,61**
CVGN	**4,36**	1,26
CVN	**3,61**	1,77
VC	0,47	**2,55**
CVG	**2,35**	0,31
VN	**1,70**	0,94
CCV	0,62	**1,55**
VG	**1,08**	0,43
CVGC	**1,20**	0,01

Quadro 4: Frequência relativa dos diferentes tipos silábicos com frequência superior a 1% presentes no *corpus* TA90PE em função da da presença/ausência de acento de palavra (num total de 16 486 sílabas tónicas e 25 359 sílabas átonas).

Quanto aos restantes tipos, importa dizer que, em geral, a ordem de frequência relativa face aos restantes tipos silábicos mantém-se inalterada nas duas posições, embora baixem de frequência face aos restantes tipos, de um modo mais significativo, as sílabas com o formato CVGN e CVG em posição átona, quando comparada com a posição tónica. Inversamente, sobem significativamente os tipos CCV e VC em posição átona, quando comparada a sua frequência de ocorrência com os outros tipos silábicos em posição tónica.

Uma exploração mais detalhada destes dados merece ser feita em trabalho futuro.

3.2. Resultados novos (2): aquisição dos tipos silábicos

Os resultados descritos acima, obtidos com a nova versão do *FreP*, foram integrados numa investigação mais vasta, onde se compararam dados de frequência dos diversos tipos silábicos na fala adulta, na fala adulta dirigida à criança e na fala da criança (Frota, Freitas, Vigário e Martins, 2005; Freitas, Frota, Vigário e Martins, 2005). Nesta investigação, mostra-se como, em relação à generalidade dos tipos silábicos, a sua ordem de emergência se correlaciona com a frequência relativa desses tipos no *input*. Todavia, existem tipos silábicos que emergem precocemente face ao esperado em função da sua frequência no discurso do adulto. A hipótese explorada é a de que o facto de esses tipos serem mais frequentes em posições proeminentes da palavra (as suas fronteiras inicial e/ou final e/ou a posição acentuada) leva a que eles se *salientem* relativamente aos restantes tipos mais frequentes e assim seja potenciado o seu mais rápido aparecimento na fala da criança.

Resultados novos (3): frequência relativa das classes de segmentos

Um outro conjunto de dados novos extraídos com o recurso à nova versão do *FreP* diz respeito à frequência relativa das grandes classes de segmentos. O Quadro 5 sintetiza o essencial dos resultados encontrados no *corpus* em estudo.

Classes de Segmentos	Totais	%
C	39 921	46%
V	41 888	48%
G	5 279	6%
Total	87 006	100%

Quadro 5: Frequência relativa das grandes classes de segmentos (C=Consoante; V=Vogal; G=Glide) no *corpus* TA90PE; valores absolutos e valores percentuais.

Estes resultados parecem-nos de grande interesse, uma vez que eles revelam que a proporção de segmentos vocálicos e de segmentos não-vocálicos é muito semelhante no Português. Embora não o façamos aqui, entendemos que este facto merece ser equacionado no contexto da discussão sobre as medidas que permitem classificar ritmicamente as línguas. Efectivamente, será interessante determinar até que ponto os resultados obtidos com uma medida deste tipo se podem correlacionar com os obtidos com o recurso à determinação do espaço (=duração do intervalo acústico) vocálico e do espaço consonântico nas línguas (cf. Ramus, Nespor e Mehler, 1999; Frota e Vigário, 2001; Frota, Vigário e Martins, 2002).

Quanto às ocorrências do que designámos *V-Stlots*, isto é os casos em que surge uma posição entre consoantes adjacentes indutoras da violação de princípios de silabificação (ver secção 2.2 acima), elas totalizam 82 casos. Em termos relativos, este valor representa 0.2% do total de sílabas/posições vocálicas do *corpus* em análise. O diminuto número de ocorrências de sequências consonânticas deste tipo no Português

(Europeu) foi já apontado e discutido em Vigário e Falé (1994), com base num estudo sobre o *corpus* do *Português Fundamental*.

Para terminar esta secção, resta-nos referir que, do total de vogais presentes no *corpus* em tratamento, 16% são nasais. Note-se que dados de frequência de vogais nasais podem ser extraídos de Viana *et al.* (1996: Quadro 4, p. 487), tendo em conta diferentes *corpora* (o que aliás é verdadeiro também para as classes de segmentos referidas anteriormente). Contudo, importa notar que, porque os valores que aí se apresentam são percentuais e relativos ao total de segmentos, não é possível aferir de um modo imediato e em valores absolutos o que representa o total de 6.7% de vogais nasais relativamente, apenas, à classe das vogais. Esta e outras operações ficam agora disponíveis ao público através da utilização desta nova versão do *FreP*.

3.4. Resultados novos (4): frequência relativa das diferentes posições do acento de palavra

Uma nova funcionalidade do *FreP* é a determinação do número de palavras com acento final, na penúltima posição e na antepenúltima. O Quadro 6 sintetiza os resultados obtidos correndo o programa sobre o *corpus* TA90PE.

Posição do acento	Totais	%
Última sílaba	2 541	21.56%
Penúltima sílaba	9 008	76.44%
Antepenúltima sílaba	235	1.99%
Total de palavras acentuadas com mais de uma sílaba	11 784	

Quadro 6: Frequência relativa das palavras em função da posição do acento de palavra no *corpus* TA90PE; valores absolutos e valores percentuais.

Os dados apresentados corroboram a intuição dos falantes sobre a localização mais frequente do acento no Português – claramente a penúltima sílaba. No mesmo sentido vão os resultados apresentados em gráfico por Viana *et al.* (1996: 494) com base noutros *corpora*.

Prevê-se dotar muito brevemente o *FreP* da capacidade de fornecer os números exactos de ocorrência de cada tipo silábico em função da posição na palavra e da presença ou ausência de acento. Assim se poderá perceber melhor como se relacionam estes dados com os princípios propostos na literatura para a colocação do acento que têm em conta a estrutura morfológica da palavra (cf. Mateus, 1983; Pereira, 1999; Mateus e Andrade, 2000; Vigário, 2003).

4. Perspectivas futuras

São duas as linhas de trabalho que pretendemos continuar a desenvolver no quadro do projecto *FreP*, uma ligada à construção da ferramenta e outra à sua exploração para o desenvolvimento do conhecimento em áreas diversas dos estudos limguísticos.

4.1. Intervenções na ferramenta

Entre as intervenções na ferramenta a fazer brevemente, conta-se a adição de um conjunto de novas funcionalidades e o completar de alguns dos paradigmas já iniciados. Pretende-se, designadamente, (i) detectar, contar e listar (o máximo de) sub-classes de segmentos; (ii) detectar, contar e listar classes de segmentos por posição na palavra, em função da presença/ausência de acento e/ou do tipo de palavra (acentuada ou clítica); (iii) detectar, contar e listar combinações de segmentos por posição na palavra, presença/ausência de acento e/ou tipo de palavra; (iv) contar a frequência de ocorrência de cada palavra e fornecer lista de frequência de palavras.

Para além disso, pretende-se dotar o *FreP* da possibilidade de editar os dados tratados pela ferramenta.

Uma tarefa de maior envergadura, também já projectada, será a de tornar possível que o *FreP* corra sobre outro tipo de escrita, em particular, sobre textos em SAMPA.

Ainda no domínio do trabalho que importa fazer muito proximamente sobre o programa, encontra-se uma avaliação sistemática e exaustiva do resultado das operações possibilitadas pela ferramenta, incluindo, para cada funcionalidade, a determinação da taxa de erro.

4.2. Linhas de trabalho a desenvolver sistematicamente

Quanto às áreas de aplicação dos dados extraídos através do *FreP*, espera-se investir num futuro próximo em duas grandes áreas fundamentais: (i) constituir informação de referência baseada na frequência, com vista à sua disponibilização para efeitos de diagnóstico, (re)avaliação e/ou terapia de fala, ou outros; (ii) explorar o uso das potencialidades do *FreP* no âmbito do ensino da língua.

Referências

Andrade, E. e A. Kihm (1988) Fonologia auto-segmental e nasais em Português. In *Actas do III Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa: APL, 51-60.

Andrade, E. e M.C. Viana (1994) Sinérese, diérese e estrutura silábica. In *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri, pp. 31-42.

Bybee, J. e P. Hopper (2001) (orgs.) *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

Beckman, M. e J. Edwards (2000) Lexical frequency effects on young children's imitative productions. In M. B. Broe e J. B. Pierrehumbert (orgs.) *Papers in Laboratory Phonology. Acquisition and the Lexicon*. Cambridge: Cambridge University Press, pp. 250-268.

Booij, G. (1995) *The Phonology of Dutch*. Oxford: Clarendon Press.

Booij, G. (2004) The morphology-phonology interface in European Portuguese. *Journal of Portuguese* Linguistics 3(1), pp. 175-182.

Demuth, K. e M. Johnson (2003) Truncation to subminimal words in Early French. *Canadian Journal of Linguistics*, 48, pp. 211-241.

Freitas, M. J., S. Frota, M. Vigário e F. Martins (2005) Efeitos prosódicos e efeitos de frequência no desenvolvimento silábico em Português Europeu. Comunicação

apresentada no *XXI Encontro Nacional da Associação Portuguesa de Linguística*, Porto, Setembro de 2005.

Frota, S., M. J. Freitas, M. Vigário e F. Martins (2005) Prosody and frequency effects on the development of syllable structure in European Portuguese. Comunicação apresentada no Simpósio *Exploring the Effects of Prosody, Morphology, Frequency and Representation on the Development of Syllable Structure in Romance Languages*, integrado no *Xth International Congress for the Study of Child Language*, Berlim, Julho de 2005.

Frota, S. e M. Vigário (2001) On the correlates of rhythmic distinctions: the European Portuguese/Brazilian Portuguese case". *Probus* 13(2), pp. 247-275.

Frota, S., M. Vigário e F. Martins (2002) Language Discrimination and Rhythm Classes: Evidence from Portuguese. In B. Bel e I. Marlien (eds.) *Speech Prosody 2002 – Proceedings of the 1st International Conference on Speech Prosody*. Aix-en-Provence: Laboratoire de Parole et Language, Université de Provence, pp. 315-318.

Jurafsky, D., A. Bell e C. Girand (2002) The Role of Lemma in Form Variation. In N. Warner e C. Gussenhoven (eds.) *Papers in Laboratory Phonolgy VII*. Cambridge: Cambridge University Press, pp. 3-34.

Mateus, M.H. (1983) O acento de palavra em Português: uma nova proposta. *Boletim de Filologia* 28: 211-229.

Mateus, M.H. e E. Andrade (2000) *The Phonology of Portuguese*. Oxford: Oxford University Press.

Pereira, I. (1999) *O Acento de Palavra em Português. Uma Análise Métrica*. Dissertação de Doutoramento, Universidade de Lisboa.

Prieto, P. (2004) Early prosodic word acquisition in Catalan. Comunicação apresentada no *Second Lisbon Meeting on Language Acquisition – with special reference to Romance Languages*, Lisboa, Junho.

Ramus, F., M. Nespor e J. Mehler (1999). Correlates of linguistic rhythm in speech. *Cognition* 73: 265-292.

Roark, B. e Demuth, K. (2000) Prosodic constraints and the learner's environment: A corpus study. In S. K. Howell, S. A. Fish e T. Keith-Lucas (orgs.), *Proceedings of the 24th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, pp. 597-608.

Thornton, A. (1996) On Some Phenomena of Prosodic Morphology in Italian: Accorciamenti, Hipocoristics and Prosodic Delemitation. *Probus* 8, pp. 81-112.

Viana, M. C., I. M.Trancoso, F. M.Silva, G. Marques, E. d'Andrade e L. C. Oliveira (1996) Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu. In *Actas do Congresso Internacional sobre o Português*, I. Duarte e I. Leiria (orgs.), vol. III. Lisboa: Colibri/APL, pp. 481-517.

Vigário, M. (2003) The Prosodic Word in European Portuguese. Berlin/New York: Mouton de Guyter.

Vigário, M., M. J. Freitas e S. Frota (no prelo) Grammar and frequency effects in the acquisition of the Prosodic Word in European Portuguese. *Language and Speech (Special Issue on the Acquisition of the Prosodic Word*, editado por Katherine Demuth).

Vigário, M., F. Martins e S. Frota (2005) Frequências no Português: a ferramenta *FreP*. In Inês Duarte e Isabel Leiria (orgs.) *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*, 897-908.

Vigário, M. e I. Falé (1994) A Sílaba no Português Fundamental: uma descrição e algumas considerações de ordem teórica. In *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri, pp. 465-477.